Bayesian Circle Segmentation with Application to DNA Copy Number Alteration Detection

Junfeng Liu[1], E. James Harner[2], Harry Yang[3]
[1] *Case Western Reserve University,* [2] *West Virginia University and*
[3] *MedImmune, Inc.*

Abstract: Several statistical approaches have been proposed to consider circumstances under which one universal distribution is not capable of fitting into the whole domain. This paper studies Bayesian detection of multiple interior epidemic/square waves in the interval domain, featured by two identical statistical distributions at both ends. We introduce a simple dimension-matching parameter proposal to implement the sampling-based posterior inference for special cases where each segmented distribution on a circle has the same set of regulating parameters. Molecular biology research reveals that, cancer progression may involve DNA copy number alteration at genome regions and connection of two biologically inactive chromosome ends results in a circle holding multiple epidemic/square waves. A slight modification of a simple novel Bayesian change point identification algorithm, random grafting-pruning Markov chain Monte Carlo (RGPMCMC), is proposed by adjusting the original change point birth/death symmetric transition probability with a differ-by-one change point number ratio. The algorithm performance is studied through simulations with connection to DNA copy number alteration detection, which promises potential application to cancer diagnosis at the genome level.

Key words: Dimension-matching, DNA copy number, Gibbs sampler, random grafting-pruning Markov chain Monte Carlo (RGPMCMC), symmetric transition.

1. Introduction

Change point models usually incorporate either a single series of observations where change points are taken as “separations” of neighboring distinct segments described by individual statistical distributions, or across multiple serials of signals where change points are taken as physical locations in a continuous one-dimensional linear space (Liu et al, 2006). From Bayesian viewpoint, the aforementioned comprehensive statistical model may be described by dimension-varying parameter θ, which takes the form of θ_K, $K = 1, 2, \ldots$, where K is the dimension and θ_K is the parameter of dimension K. The prior distribution $\pi(\theta)$ could be written as $\sum_{K\in\mathcal{N}}\pi(\theta_K\,|\,K)\pi(K)$, where $\mathcal{N}$ represents the positive integer set, $\pi(K)$ is the mixture probability for dimension K and $\pi(\theta_K\,|\,K)$ is the individual prior distribution within the K-dimensional space. For regularity, we assume $\sum_{K\in\mathcal{N}}\pi(K){=}1$ and $\int_{\Theta_K}\pi(\theta_K\,|\,K)d\theta_K{=}1$ for each $K\in\mathcal{N}$. Since $\pi(K,\theta_K)$ equals $\pi(K)\pi(\theta_K\,|\,K)$, the posterior distribution of (K,θ_K) is

$$\pi(K,\theta_K\,|\,X)=\frac{f(X\,|\,\theta_K)\pi(\theta_K\,|\,K)\pi(K)}{\sum_{K\in\mathcal{N}}(\int_{\Theta_K}f(X\,|\,\theta_K)\pi(\theta_K\,|\,K)d\theta_K)\pi(K)}, \tag{1.1}$$

where the denominator is the normalization constant not needed for posterior sampling. $(K,\,\theta_K)$ is a realization of dimension-varying θ, where only the second component θ_K is sufficient to represent a possible value of θ and the first component K only helps to induce the prior density as well as the posterior density in (1.1). Sampling-based approaches for change point detection in the literature are the reversible jump Markov chain Monte Carlo (RJMCMC) by Green (1995), the continuous time birth/death process MCMC by Stephens (2000), the product partition model based Bayesian algorithm by Loschi, Cruz and Arellano-Valle (2005) which stems from Yao (1984) and Barry and Hartigan (1992, 1993), non-MCMC based recursive sampling algorithm by Fearnhead (2005) and others. Denison et al (2002) observe that, the partitioning and wrapping up the segments leads to an exponentially increasing computational cost as the discrete model space grows. From frequentist perspective, Sen and Srivastava (1975) tested whether the means of independent sequential random variables are the same or there is a shift after some point; Olshen et al (2004) developed a frequentist sequential circular binary segmentation (CBS) algorithm to detect multiple interior waves in a linear domain, which is essentially a circular domain by connecting two ends with identical statistical properties. Circular domains are broadly existent in the real world, e.g., subway route, electric current in wire loop exposed under magnetic field, remanent magnetization distribution in rock samples (Fisher and Lee, 1986) and certain circular DNA molecules (Stanfield and Lengyel 1979) among others. Specifically, Mailfert and Vincent-Viry (2001) studied generating uniform force field in circular domain for energy storage, and Fouts et al (1978) experimentally identified the relative change sites of cleavage sensitivity to enzymes *EcoRI* and *HindIII* along *C.acanthocephali* circular kDNA molecule. Compared with linear domain segmentation where the first segment always starts from the left domain end point and the last segment always ends at the right domain end point, the circular domain segmentation is to be implemented without explicit end points. We may cut the circle open at any preselected point on the circle, expand it into a linear domain and apply the linear segmentation introduced by Liu et al (2006). However, this may complicate the work on two end segments since they almost surely belong to the same segment (distribution) without the need for redundant regulating parameters. In other words, any preselected reference point almost surely falls into the interior of certain segment, other than onto certain segment boundary. The present work proposes a simple Bayesian version of multiple interior epidemic/square wave (change point) detection, where each segmented distribution has the same set of regulating parameters. This is a slight modification of the random grafting-pruning Markov chain Monte Carlo (RGPMCMC) algorithm introduced by Liu et al (2006). One of the promising applications is for biomarker identification in modern bioinformatics research.

The organization of this article is as follows: Section 2 introduces the background of Bayesian dimension-matching and proposes our simple sampling algorithm; Section 3 tests the algorithm by several simulations under diverse configurations with application to DNA copy number alteration detection; Section 4 concludes with discussion.

2. Random Grafting-pruning Markov Chain Monte Carlo (RGPMCMC)

We introduce a common dimension-varying scenario which we aim to work on: Several disjoint segments exist on a circle without gaps, each segment has an individual set of observations regulated by segment-specific means and variances, and the interested observations are indexed by the locations along the circle. Our objective is to identify the segment boundaries as well as the regulating means and variances. We use a graphical model to demonstrate the data and our proposed algorithm. For certain $K \in \mathcal{N}$ and a circle with circumference L, there are K change points along it to form K segments C_k, $k = 1, \ldots, K$. After preselecting an arbitrary reference origin on the circle, say O, the clockwise distance away from O to any point on the circle is taken as this point's location such that any location is between 0 and circle circumference L, and by this way we make all locations comparable in terms of magnitude. For each $k \in [1, K]$, C_k spans from location t_k to t_{k+1} and holds n_k data point pairs (x_{kj}, y_{kj}), $j = 1, \ldots, n_k$, where x_{kj} is the location along the circle such that $t_k \leq x_{kj} \leq t_{k+1}$, i.e., $x_{kj} \in C_k$, $1 \leq j \leq n_k$, and y_{kj} is our interested measurement indexed by location x_{kj}, which can be visualized as a centrifugal/zero/centripetal distance away from the circle boundary if y_{kj} is positive/zero/negative. t_{K+1} equals t_1 for notational convenience. Note that, the reference origin is only used to assign change point location vector $\tilde{T}_K = (t_1, t_2, \ldots, t_K)$ given any $K \in \mathcal{N}$ and the fulfillment of the sampling algorithm does not require a specific origin or circle circumference scale (the unit of L), although the prior density may depend on L value (details later in parameter sampling process). Assuming a normal distribution $N(\mu_k, \sigma_k^2)$ for all y's (directional distances away from circle boundary) whose locations (x's) are within C_k , $k = 1, \dots, K$ and $K \in \mathcal{N}$, we are interested in estimating: the most likely change point number K, change point location (segment starting location) vector $\tilde{T}_K = (t_1, t_2, \dots, t_K)$ (segmenting x's) as well as parameter sets (μ_k, σ_k^2) for these K normal distributions. As for change point birth/death, we propose a comprehensive stochastic parameter sampling process where four move types are considered to achieve the desired posterior sampling from (1.1). The motivation is described as follows: without change point birth/death, we are working with ordinary posterior sampling under same dimension, where some routine procedure follows, say applying conjugate normal or inverse-gamma posterior distributions to update μ_k and σ_k^2, $k = 1, \dots, K$, or Metropolis-Hastings algorithm for change point location $\tilde{T}_K$ updating, the other necessary procedure is to realize change point birth/death, which results in dimension-changing. Once these tasks are fulfilled, a dimension-matching parameter sampling process will be mobilized. Specifically, we use the same notations for these four move types (H, P, $+$, $-$) as suggested by Green (1995), where "H" represents the proposal for all those regulating parameters (μ_k, σ_k^2, $k = 1, \dots, K$) other than change point location $\tilde{T}_K$; "P" represents change point location $\tilde{T}_K$ proposal without change point birth/death; "$+$" represents change point birth proposal (new t_*, μ_* and σ_*^2); "$-$" represents change point death proposal (deleting certain t_*, μ_* and σ_*^2). We specify (H, P, $+$, $-$) probabilities to be ($\pi(H)$, $\pi(P)$, $\pi(+)$, $\pi(-)$) with summation of one and give transdimensional change relatively higher probabilities, say 0.9 (details follow in the parameter sampling process). For convenience, we apply N-Inv-$\chi^2(\mu_0,\sigma_0^2/\kappa_0;\nu_0,\sigma_0^2)$ conjugate prior (Gelman et al, 1995) to parameter pair (μ,σ^2), where $\mu \,|\, \sigma^2 \sim \mathrm{N}(\mu_0,\sigma^2/\kappa_0)$ and $\sigma^2 \sim \text{Inv-}\chi^2(\nu_0,\sigma_0^2)$, corresponding to the following joint prior density

$$\pi(\mu, \sigma^2) \propto \sigma^{-1}(\sigma^2)^{-(\nu_0/2+1)} \exp(-\frac{1}{2\sigma^2}[\nu_0\sigma_0^2 + \kappa_0(\mu_0 - \mu)^2]). \qquad (2.1)$$

We let $y = (y_1, y_2, \dots, y_n)$ represent the observed independent random variables coming from $N(\mu, \sigma^2)$ with density function $\prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(y_i - \mu)^2)$. The joint posterior density is thus

$$\begin{aligned} \pi(\mu, \sigma^2 \,|\, y) &\propto \sigma^{-1}(\sigma^2)^{-(\nu_0/2+1)} \exp(-\frac{1}{2\sigma^2}[\nu_0\sigma_0^2 + \kappa_0(\mu - \mu_0)^2]) \\ &\qquad \times (\sigma^2)^{-n/2} \exp(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2]) \\ &\sim \text{N-Inv-}\chi^2(\mu_n, \sigma_n^2/\kappa_n; \nu_n, \sigma_n^2), \end{aligned}$$

where, s^2 and $\bar{y}$ are the sample variance and sample mean of y respectively. It can be shown that,

$$\begin{aligned}
\mu_n &= \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y} \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_0 + n \\
\nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0+n}(\bar{y}-\mu_0)^2.
\end{aligned}$$

It is known that the marginal posterior distribution of σ^2 is scaled Inv-χ^2 with

$$\sigma^2 \,|\, y \sim \text{Inv-}\chi^2(\nu_n, \sigma_n^2) \tag{2.2}$$

and the conditional posterior distribution of μ given σ^2 is

$$\mu \,|\, \sigma^2, y \sim \mathrm{N}(\mu_n, \sigma^2/\kappa_n) = \mathrm{N}\left(\frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_0+n}, \frac{\sigma^2}{\kappa_0+n}\right). \tag{2.3}$$

Thus the joint posterior sampling of (μ,σ^2) within each segment is to be implemented through (2.2) followed by (2.3). Let $\tilde{\mu}_K$ and $\tilde{\sigma}_K^2$ represent all (μ_k, σ_k^2) pairs, $\boldsymbol{X}$ and $\boldsymbol{Y}$ represent all (x_{kj},y_{kj}) pairs, $1 \le j \le n_k$, $1 \le k \le K$, we obtain the following likelihood function for the model described at the beginning of this section

$$f(\boldsymbol{X}, \boldsymbol{Y} \,|\, K, \tilde{T}_K, \tilde{\mu}_K, \tilde{\sigma}_K^2) = \prod_{k=1}^{K} \prod_{1\le j\le n_k} \frac{1}{\sqrt{2\pi}\sigma_k} \exp(-\frac{1}{2\sigma^2}(y_{kj}-\mu_k)^2), \tag{2.4}$$

and the posterior distribution is simply

$$\begin{aligned}
& f(K, \tilde{T}_K, \tilde{\mu}_K, \tilde{\sigma}_K^2 \,|\, \boldsymbol{X}, \boldsymbol{Y}) \\
\propto\ & f(\boldsymbol{X}, \boldsymbol{Y} \,|\, K, \tilde{T}_K, \tilde{\mu}_K, \tilde{\sigma}_K^2)\pi(K)\pi(\tilde{T}_K \,|\, K)[\prod_{k=1}^{K} \pi(\mu_k, \sigma_k^2)] \\
=\ & \pi(K)\pi(\tilde{T}_K \,|\, K) \\
& \times \prod_{k=1}^{K} (\prod_{1\le j\le n_k} \frac{1}{\sqrt{2\pi}\sigma_k} \exp(-\frac{1}{2\sigma_k^2}(y_{kj}-\mu_k)^2))\pi(\mu_k, \sigma_k^2),
\end{aligned} \tag{2.5}$$

where, locations $\boldsymbol{X}$ only help to construct the likelihood function of $\boldsymbol{Y}$ and the posterior distribution only incorporates $\boldsymbol{Y}$ and our interested parameters. Note that, the ($\tilde{T}_K$,$\tilde{\mu}_K$,$\tilde{\sigma}_K^2$) is the θ_K in (1.1) for each $K \in \mathcal{N}$. If we assume $\pi(K)$ follows a truncated Poisson or uniform discrete prior between two positive integers, then K belongs to a subset of $\mathcal{N}$ *a priori* and each $\pi(K)$ in (1.1) is zero for those K's which are not within this subset; we assume $\pi(\mu_k, \sigma_k^2)$ follows joint conjugate prior (2.1). As suggested by Green (1995), to avoid too short segments we assume $\pi(\tilde{T}_K \,|\, K)$ is the distribution of even order statistics from uniform random variables along a circle with density

$$\pi(\tilde{T}_K \,|\, K) = \frac{(2K-1)!}{L^{2K}} \prod_{k=1}^{K} (t_{k+1} - t_k),$$

where $t_1 = t_{K+1}$.

Proposition 1 On a circle with circumference L, the K dimensional $(m+1)$-th ordered uniform random variables $(t_1, t_2, \ldots, t_K)$=$(X_{(0)}, X_{(m+1)}, X_{(2m+2)}, \ldots, X_{(Km+K)})$ has the following distribution density

$$\pi(t_1, t_2, \ldots, t_K) = \frac{(K(m+1)-1)!}{L^{K(m+1)}} \prod_{k=1}^{K} (t_{k+1} - t_k),$$

where $t_1 = t_{K+1}$.

Proof.

Given random variable indexes and equal number (m) of random variables between every $(m+1)$-th change point,

$$\begin{aligned} &Pr(T_1 \in (t_1, t_1 + \Delta_1), T_2 \in (t_2, t_2 + \Delta_2), \ldots, T_K \in (t_K, t_K + \Delta_K)) \\ = &\prod_{k=1}^{K} [(\frac{\Delta_k}{L})(\frac{t_{k+1} - t_k}{L})^m], \end{aligned}$$

the final result is observed by letting Δ_k $(k = 1, 2, \ldots, K)$ approaching zero and confounding the K random variable indexes for these $(m+1)$-th ordered random variables by exhaustive permutation.

Parameter Sampling Process

1). First we choose one of these four move types based on move type probabilities ($\pi(H)$, $\pi(P)$, $\pi(+)$, $\pi(-)$), where $\pi(+)$ is taken to be equal to $\pi(-)$, say 0.45.

2). For "H" move type, we refer to the conjugate posterior distributions from equations (2.2) and (2.3). $\bar{y}_k = \frac{1}{n_k} \sum_{1 \le j \le n_k} y_{kj}$ and $s_k^2 = \frac{1}{n_k - 1} \sum_{1 \le j \le n_k} (y_{kj} -$

$\bar{y}_k)^2$. σ_k^2 is to be sampled from its marginal posterior distribution Inv-$\chi^2(\nu_k'$, $\eta_k')$, where

$$\nu_k' = \nu_0 + n_k \quad \text{and} \quad \eta_k' = \frac{\sigma_0^2\nu_0}{\nu_n} + \frac{(n_k-1)s_k^2}{\nu_n} + \frac{\kappa_0 n_k}{\nu_n(\kappa_0+n_k)}(\bar{y}_k - \mu_0)^2,$$

then μ_k is to be sampled from its posterior conditional distribution N$(\zeta_k$,$\phi_k^2)$ given σ_k^2, where $\zeta_k = (\kappa_0\mu_0 + n_k\bar{y}_k)/(\kappa_0+n_k)$ and $\phi_k^2 = \sigma_k^2/(\kappa_0+n_k)$.

3). For "P" move type, we randomly sample a change point separating two neighboring segments, say t_*, with equal probability $1/K_{old}$, and a location is uniformly randomly selected within these two fused segments for a new substituting candidate change point to replace the current separation point t_*. The other parameters associated with this change point mutation proposal is kept unchanged. This is a symmetric transition (Proposition 3) leading to following acceptance probability in the Metropolis-Hastings algorithm within Gibbs sampler

$$\min\left\{1,\ \frac{f(K,\tilde{T}_K,\tilde{\mu}_K,\tilde{\sigma}_K[\text{after }] \mid \boldsymbol{Y})}{f(K,\tilde{T}_K,\tilde{\mu}_K,\tilde{\sigma}_K[\text{before CPMP}] \mid \boldsymbol{Y})}\right\},$$

where CPMP stands for "change point mutation proposal" for short.

4). For "+" move type, if $K_{old} = K_{max}$, we go to 1) since the maximum threshold is reached; if $K_{old} < K_{max}$, we randomly sample one of the K_{old} segments formed by current K_{old} change points with equal probability $1/K_{old}$, say C_j. Within this sampled segment, we uniformly sample a candidate change point birth at location t_*, a new segment is thus embedded into current segments C_j and C_{j+1}, then we propose non-change point parameters accompanying this change point location candidate by $\mu_* = \log(\frac{U_1}{1-U_1})(\mu_j + \mu_{j+1})/2$ and $\sigma_*^2 = (\frac{U_2}{1-U_2})\sigma_j\sigma_{j+1}$, where U_1 and U_2 $\sim U[0,1]$ independently. Based on the observations in Liu et al (2006), the segment birth proposal involved in "+" move type, along with the segment death proposal in the following "$-$" move type, constructs a symmetric transition in the one-dimensioanl infinitesimal space which holds the potential change points. Proposition 2 observes that, the acceptance probability in the Metropolis-Hastings algorithm within Gibbs sampler for circular domain is simply

$$\min\left\{1,\ \frac{f(K,\tilde{T}_K,\tilde{\mu}_K,\tilde{\sigma}_K[\text{after CPDP}] \mid \boldsymbol{Y})}{f(K,\tilde{T}_K,\tilde{\mu}_K,\tilde{\sigma}_K[\text{before CPDP}] \mid \boldsymbol{Y})} \times \frac{K_{old}}{K_{old}+1}\right\},$$

where CPDP stands for "change point death proposal" for short.

5). For "−" move type, if $K_{old} = K_{min}$, we go to 1) since the minimum threshold is reached; if $K_{old} > K_{min}$, we randomly sample one from current K_{old} change points with equal probability $1/K_{old}$, say at location t_*, to delete. Then we simply delete the associated μ_* and σ_*^2 for change point at t_*, and the likelihood function is to be computed accordingly after deletion. The acceptance probability in the Metropolis-Hastings algorithm within Gibbs sampler is simply (Proposition 2)

$$\min\left\{1,\ \frac{f(K,\tilde{T}_K,\tilde{\mu}_K,\tilde{\sigma}_K[\text{after CPDP}]\mid \boldsymbol{Y})}{f(K,\tilde{T}_K,\tilde{\mu}_K,\tilde{\sigma}_K[\text{before CPDP}]\mid \boldsymbol{Y})}\times\frac{K_{old}}{K_{old}-1}\right\}.$$

We may propose a more general framework:

- ◇ Change point location birth: after selecting segment C_j to work on, we let

$$t_* = g_{t_*}(U_1; t_j, t_{j+1}) \in (t_j, t_{j+1}),$$

where $g_{t_*}(U_1; t_j, t_{j+1})$ is a one-to-one mapping from random variable U_1 to change point location t_* given allowable domain $(t_j,\ t_{j+1})$. We take $g_{t_*}(U_1; t_j, t_{j+1})$ to be $(t_j + t_{j+1}g_1(U_1))/(1 + g_1(U_1))$, i.e., $(t_* - t_j)/(t_{j+1} - t_*){=}g_1(U_1)$, where $g_1(\cdot)$ is any monotonic function with domain (0,1) and range (0,∞), and $U_1 \sim U[0,1]$. It can be seen that t_* is a monotonically increasing function of U_1. We simply use $g_1(u) = u/(1-u)$, thus $t_* = t_j + (t_{j+1} - t_j)U_1$, a uniform random variable $\in(t_j,\ t_{j+1})$;

- ◇ Lifted normal distribution mean μ_*: we proceed by letting

$$\mu_* = g_{\mu_*}(U_2; \mu_j, \mu_{j+1}),$$

where $g_{\mu_*}(U_2; \mu_j, \mu_{j+1})$ is a one-to-one mapping from random variable U_2 to lifted normal mean μ_* given neighboring μ_j and μ_{j+1}. We take $g_{\mu_*}(U_2; \mu_j, \mu_{j+1})$ to be $g_2(U_2)(\mu_j + \mu_{j+1})/2$, where $g_2(\cdot)$ is any monotonic function with domain (0,1) and range (-∞,+∞), and U_2 $\sim U[0,1]$. It can be seen that μ_* is a monotonically increasing function of U_2. We simply use $g_2(u) = \log(u/(1-u))$ in our proposal;

- ◇ Lifted normal distribution variation σ_*: we proceed by letting

$$\sigma_*^2 = g_{\sigma_*^2}(U_3; \sigma_j^2, \sigma_{j+1}^2),$$

where $g_{\sigma_*^2}(U_3;\sigma_j^2,\sigma_{j+1}^2)$ is a one-to-one mapping from random variable U_3 to lifted normal variance σ_*^2 given current neighboring σ_j^2 and σ_{j+1}^2. We take $g_{\sigma_*^2}(U_3;\sigma_j^2,\sigma_{j+1}^2)$ to be $(\sigma_j^2\sigma_{j+1}^2)^{1/2}g_3(U_3)$, i.e., $(\sigma_*/\sigma_j)/(\sigma_{j+1}/\sigma_*)$=$g_3(U_3)$, where $g_3(\cdot)$ is any monotonic function with domain (0,1) and range (0,∞), and $U_3 \sim U[0,1]$. It can be seen that σ_*^2 is a monotonically increasing function of U_3. We simply use $g_3(u) = u/(1-u)$, thus $\sigma_*^2 = (\sigma_j^2\sigma_{j+1}^2)^{1/2}U_3/(1-U_3)$.

We first briefly describe a simple case where only change point locations are considered. Considering measure based probability and assuming segment C_j is randomly selected for change point (with index $*$) birth proposal, i.e., the new candidate segment starting location $t_* \in(t_j,t_{j+1})$=C_j, and A is any Borel measurable set within segment C_j. The change point (segment) birth proposal probability given current K_{old} segments is the product of $1/K_{\text{old}}$ and A/(segment C_j length), where the former one is the probability of selecting segment C_j out of current K_{old} ones, the latter one is the probability of landing in set A conditional on segment C_j, i.e.,

Proposition 2 Under stochastic "+/−" move type in the parameter sampling process, birth/death proposal has transition probability only proportional to the ratio of two differ-by-one segment numbers.

$$\frac{1}{K_{\text{old}}} \times \frac{A}{\text{segment } C_j \text{ length}}.$$

Suppose we have K_{old}+1 available segments after presumed birth proposal and consider the exact reverse (change point/segment death) transition probability by integrating the current state (preceding t_*) over A within segment C_j. The change point (segment) death proposal probability given current K_{old}+1 segments is the product of A/(segment C_j length) and 1/(K_{old}+1), where the former one is the natural probability measure of Borel set A conditional on segment C_j , the latter one is the probability of selecting change point starting at t_* out of current K_{old}+1 ones, i.e.,

$$\frac{A}{\text{segment } C_j \text{ length}} \times \frac{1}{K_{\text{old}}+1}.$$

The proof for only change point birth/death ends. We now describe the statistical properties for the comprehensive parameter sampling process arising from four move types (H, P, +, −). The segment (change point) birth/death proposal process takes account of current change point density information, e.g., dense change points attract more attention in view of the fact that, we work on segment indexes with equal weights, other than segment length-based selection. We essentially map the individual parameter set (t_*, μ_*, σ_*^2) back into the generating spaces of variables U_1, U_2 and U_3. See "+" move type and the general parameter proposal framework after parameter sampling process. The probability is only in the sense of latter spaces. The product probability of certain measurable balls around the observed parameter set

$$B(t_*) \times B(\mu_*) \times B(\sigma_*^2), \tag{2.6}$$

is corresponding to

$$B'(U_1) \times B'(U_2) \times B'(U_3), \tag{2.7}$$

where the $B'(\cdot)$s are simply the mapped generating sets for observed parameter sets $B(\cdot)$s. For change point birth, $1/K_{old}$ is the probability of randomly selecting a segment $*$ out of current K_{old} ones, out of which a candidate change point grows up at randomly chosen location, say t_*, in terms of Borel set (2.6) specified by associated parameter set (t_*,μ_*,σ_*^2). We essentially work on two parts of interested parameters: the segment index $*$ (I) along with the describing set (t_*,μ_*,σ_*^2) (D). The "+" move type realizes "I "and "D" parts sequentially while independently. We may imagine trying proposing the reverse process (change point death) by tracing the presumed change point birth proposal: a current "I", say $*$, is randomly selected with equal probability $1/K_{old}$ with available describing parameter set "D", say (t_*,μ_*,σ_*^2). The presumed birth proposal for this very change point is to be replayed independently (we are looking at the combined two segments as one possible original segment for change point birth location proposal within this selected segment, see the move type "+") within this selected segment. Any non-exact overlap with the original change point describing "D" parameters (within the same segment) represents an ineffective change point death proposal (no proposal for Metropolis-Hastings algorithm) and only an exact overlap represents an effective change point death proposal, with an obvious identical probability as random change point birth within this very segment. This equally probable reverse events within the same segment (original single segment for change point birth proposal or two joined neighboring segments for change point death proposal) simply quantitatively verifies our within-segment symmetric transition process between change point birth and death in a probabilistic manner. The ineffective change point death proposal (no proposal) leads to proposal-freezing Metropolis-Hasting algorithm based Markov chain, which is probabilistically useless for Metropolis-Hastings algorithm based inference. Conditioning on choosing one change point out of current K ones separating two neighboring segments (combined segments) on the circle, we imagine equal E-partition on all such combined segment pairs. Due to the aforementioned change point death proposal by effective exact overlap with a presumed change point birth proposal, on the circular domain the probability of realizing a specific birth for this change point is $\pi(+)(\frac{1}{K-1})(\frac{1}{E})$, the probability of realizing a specific death for this change point is $\pi(-)(\frac{1}{K})(\frac{1}{E})$, the probability of no action (proposal freezing) is $\pi(-)(\frac{1}{K})(\frac{E-1}{E})$. The aforementioned parameter proposal process statistically discards unnecessary proposal freezing for efficient Markov chain based inference. Random grafting-pruning Markov chain Monte Carlo (RGPMCMC) arises from the change point birth/death proposal which acts like randomly grafting or pruning a plant: after a random selection of the branch interval or branch, change point birth/death proposal physically takes place along the plant stem (circle), the set of sub-branches, i.e., the describing parameter set (including change point birth/segment starting location and regulating parameters for individual normal distribution), are lifted independently for change point birth proposal, or deleted for change point death proposal. The branch (segment) birth and death proposal should be committed alternatively in a probabilistic manner. If we happen to choose to randomly delete one branch, then we may randomly add this very branch in the same place in the preceding birth proposal; on the other hand, if we happen to choose to randomly add one branch, then we may randomly delete this very branch in the succeeding death proposal, the equal probable branch birth/death is realized physically (under differ-by-one branch number ratio adjustment). The information balance could also be intuitively justified in this way: among well established segments, it is equally difficult to add another segment into any existent segment, or to fuse any two neighboring segments which are already well established. The grafting step (move type "+") is quite noninformative by essentially requiring randomly harvesting a branch from the garden, which could be done by designing any convenient one-to-one mapping proposal functions ((2.6) and (2.7)); the pruning step (move type "−") is also trivial by just randomly deleting one branch. We simplifies the acceptance rate in Metropolis-Hastings algorithm as only a local adjustment, e.g., for segment birth we only reassign the y's in one original segment into two split segments (a new segment starting location comes into one of current segments), incorporate an additional prior set for new regulating parameters (t_*,μ_* and σ_*^2) and reconstruct the likelihood function due to this new segment, and vice versa for segment death.

Proposition 3 Random single change point location mutation proposal within combined segments as described in move type "P" is a symmetric transition.

Proof.

Assume two non-zero measurable sets dt and dt' are located within the two fused segments with length C, we have

$$P(dt \to dt') = \int_{dt} P(t \to dt') = \frac{dt'}{C}\frac{dt}{C} = \frac{dt}{C}\frac{dt'}{C} = \int_{dt'} P(t' \to dt) = P(dt' \to dt).$$

The "+/−" move type realizes one type of symmetric transition (equally probable change point birth and death) after segment number ratio adjustment; "P" move type realizes another type of symmetric transition (equally probable location selection within two combined segments).

Proposition 4 This Markov chain is irreducible, aperiodic.

Interpretation. In view of detailed balance proposal, there is a positive probability that the chain lies in any small neighborhood after one sampling iteration to meet the aperiodicity; the chain can move from any value to any other value in steps of one at a time to establish the irreducibility.

3. Simulation Study

We use the same notations from Section 2: L is the circle circumference, K is the number of change points along it, the K segments are C_k, $k = 1, \ldots, K$. The reference location origin is specified at an arbitrary point with nominal zero value. Segment C_k ($[t_k,\, t_{k+1}]$) holds n_k data pairs (x_{kj}, y_{kj}), $j = 1, \ldots, n_k$, where x_{kj} is the clockwise location along the circle from certain reference origin and y_{kj} is the measurement (represented by centrifugal/centripetal distance away from circle boundary). All y's in segment C_k follow a normal distribution $N(\mu_k, \sigma_k^2)$, $k = 1, \ldots, K$. We are interested in: change point number K, change point location vector $\tilde{T}_K = (t_1, t_2, \ldots, t_K)$ and parameter sets (μ_k, σ_k^2), $k = 1, \ldots, K$ for these K normal distributions. We take ($\pi(H)$=0.05, $\pi(P)$=0.05, $\pi(+)$=0.45, $\pi(-)$=0.45) to be four move type probabilities. The prior for segment number is $K \sim U[1, 100]$, and 51 is taken as starting change point number for iteration and equal circular 51-partition as initial change points. We assign L=10 for our simulation and computations. The results are given in Figures 1 and 2.

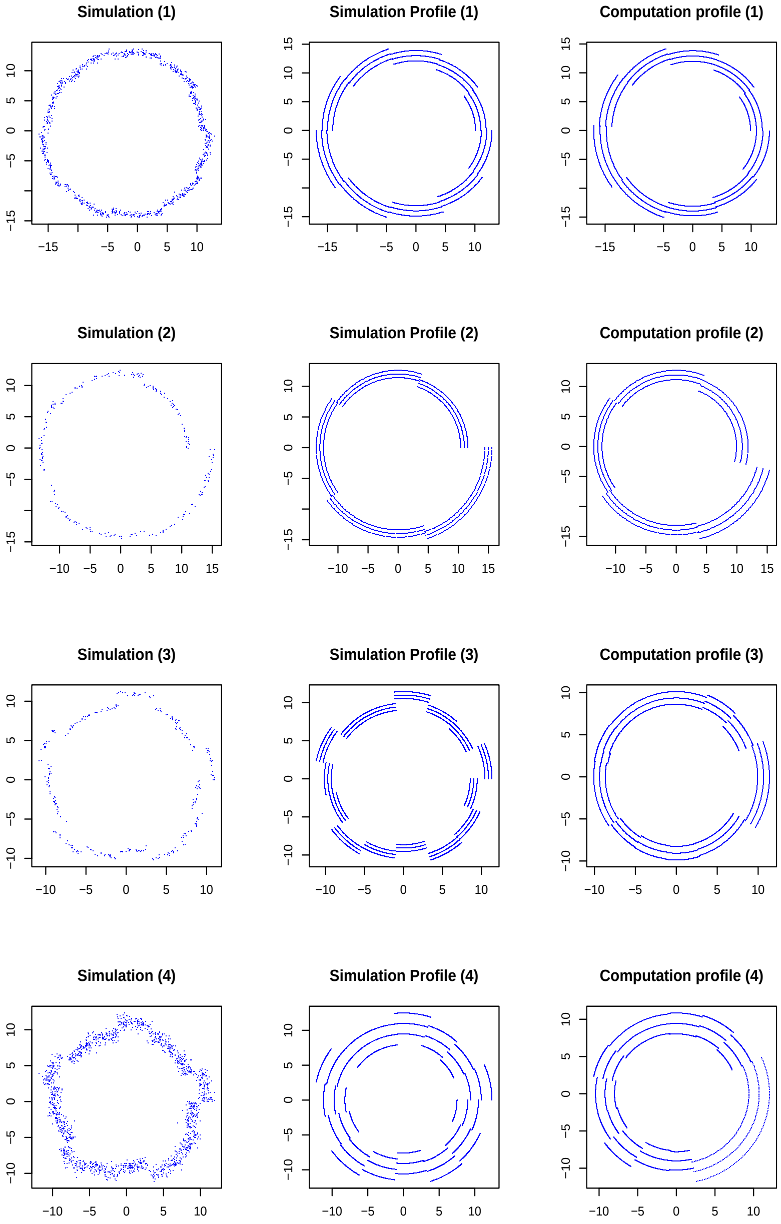

Figure 1: Simulation profiles and estimated profiles. (constructed from segment partition, segment means and segment variations).

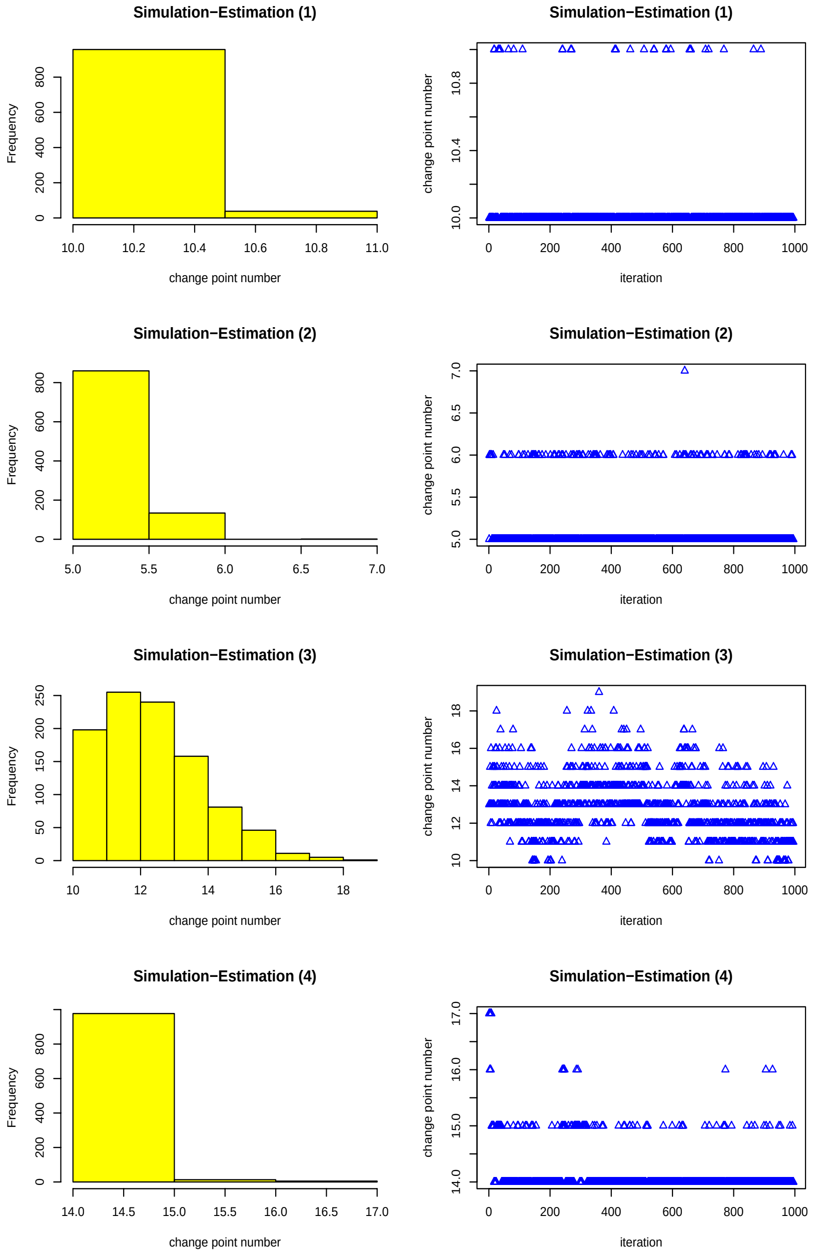

Figure 2: Posterior segment number distribution and iteration

Simulation 1.

- ◇ True parameters: $K = 10$, $n_k = 100$, $t_k = (k-1)\frac{L}{K}$, $\mu_k = k, (k = 1, \ldots, 6)$ and $\mu_k = 12 - k, (k = 7, \ldots, K)$, $\sigma_k = 0.30$;
- ◇ Priors: $\kappa_0 = 1.0$, $\nu_0 = 6.0$, $\sigma_0^2 = \frac{1}{3}$, $\mu_0 = 3.0$;
- ◇ Observations: The change points are completely identified and the estimation profile is consistent with the simulation profile.

Simulation 2.

- ◇ True parameters: $K = 5$, $n_k = 50$, $t_k = (k-1)\frac{L}{K}$, $\mu_k = k, (k = 1, \ldots, K)$, $\sigma_k = 0.20$;
- ◇ Priors: $\kappa_0 = 1.0$, $\nu_0 = 6.0$, $\sigma_0^2 = \frac{1}{3}$, $\mu_0 = 3.0$;
- ◇ Observations: Under smaller n_k the change points are completely identified and the estimation profile is wider than the simulation profile, indicating potential sensitivity to dispersion prior in this specific scenario.

Simulation 3.

- ◇ True parameters: $K = 15$, $n_k = 20$, $t_k = (k-1)\frac{L}{K}$, $\mu_k = \sin(2k), (k = 1, \ldots, K)$, $\sigma_k = 0.15$;
- ◇ Priors: $\kappa_0 = 1.0$, $\nu_0 = 6.0$, $\sigma_0^2 = \frac{1}{3}$, $\mu_0 = 0.5$;
- ◇ Observations: Under smaller μ_k variation, n_k and σ_k, the change point number follows a posterior distribution with multiple non-ignorable probabilities around change point number 11 and the estimated profile is wider than the simulation profile indicating potential sensitivity to dispersion prior in this specific scenario.

Simulation 4.

- ◇ True parameters: $K = 15$, $n_k = 100$, $t_k = (k-1)\frac{L}{K}$, $\mu_k = \sin(2k), (k = 1, \ldots, K)$, $\sigma_k = 0.50$;
- ◇ Priors: $\kappa_0 = 1.0$, $\nu_0 = 6.0$, $\sigma_0^2 = \frac{1}{3}$, $\mu_0 = 0.5$;
- ◇ Observations: Under larger n_k and larger σ_k (in contrast to simulation 3) 14 out of 15 change points are identified and the estimation profile is wider than the simulation profile, another minor posterior change point number is 15 representing complete recovery with a smaller probability.

We find that, when the segment-specific normal means are substantially different from neighbors (simulations 1 and 2), or segment-specific normal variations are small, the sampler recovers segmentation information very well; when more data points fall into each segment (simulations 1 and 4), the sampler will also identify segmentation satisfactorily. Intuitively, these two effects make the change points more visually clustered. We hypothesize that, the variation effect dominates the number of data points effect by referring to normal likelihood function (square effect vs. linear effect), and this was verified in Olshen et al (2004). The sampler locates change point number within seconds in an efficient manner and intensive local mutation pinpoints accurate change point locations around segment boundaries.

At quantitative biology research, except for enzyme sensitivity change region detection in circular kDNA molecule (Fouts et al, 1978), we now highlight an important application of our proposed algorithm to modern bioinformatics with connection to disease diagnostics at the genome level. The comparative genomic hybridization (CGH) introduced by Kallioniemi et al (1992) offers a DNA sequence copy number map along the entire genome. The test DNA and normal reference DNA are differentially labeled and hybridized simultaneously to normal chromosome spreads. Regions of gain or loss of DNA sequences, are seen as changes in the ratio of the two color intensities and change regions were identified for tumor DNA as disease biomarkers. Lucito et al (2003) developed a CGH-based large-scale ROMA (representational oligonucleotide microarray analysis) for genomic aberration detection in cancer and normal humans. They hybridized designed oligonucleotide probes from human genome sequence with "representations" from cancer and normal cells, and detected genome regions with altered copy number. ROMA could identify variation between cancer and normal genomes for potential genome-wide disease marker discovery. Some statistical tools were developed recently in the frequentist framework: Huang et al (2005) applied penalized least square estimate with consideration of spatial dependence, Lai and Zhao (2005) applied bootstrapped one-sample t-test and the false discovery rate control to identify chromosomal alteration regions, and Olshen et al (2004) developed sequential circular binary segmentation (CBS) algorithm to identify DNA copy number alterations among others. These algorithms need tuning parameters in order to obtain biologically consistent results. We apply the same simulation study from Section 5 in Olshen et al (2004) to test the performance of our Bayesian algorithm, where the prior specification takes the role of tuning parameters required by frequentist methods. The data are simulated based on the CBS fit to chromosome 11 of a real ROMA (Lucito et al, 2003) breast cancer experiment. There are 497 DNA copy number markers in chromosome 11 with six change points estimated at 137, 224, 241, 298, 307, 331 and the average log-ratios of intensities within segments are given in the following table.

Segment C_k	[1,137]	[138,224]	[225,241]	[242,298]
Mean $f(k)$	-0.18	0.08	1.07	-0.53
Segment C_k	[299,307]	[308,331]	[332,497]	
Mean $f(k)$	0.16	-0.69	-0.16	

In view of all segmentation means and the close mean values at two ends, it is reasonable to form a circle by connecting two ends into a joint segment, and this assumption is not casual data manipulation since it is biologically reasonable to assume that, DNA copy number changes are owing to long term evolution taking place in sporadic locations between two inactive chromosome ends. Massive biological data sets indicate this pattern. The data are generated using the model

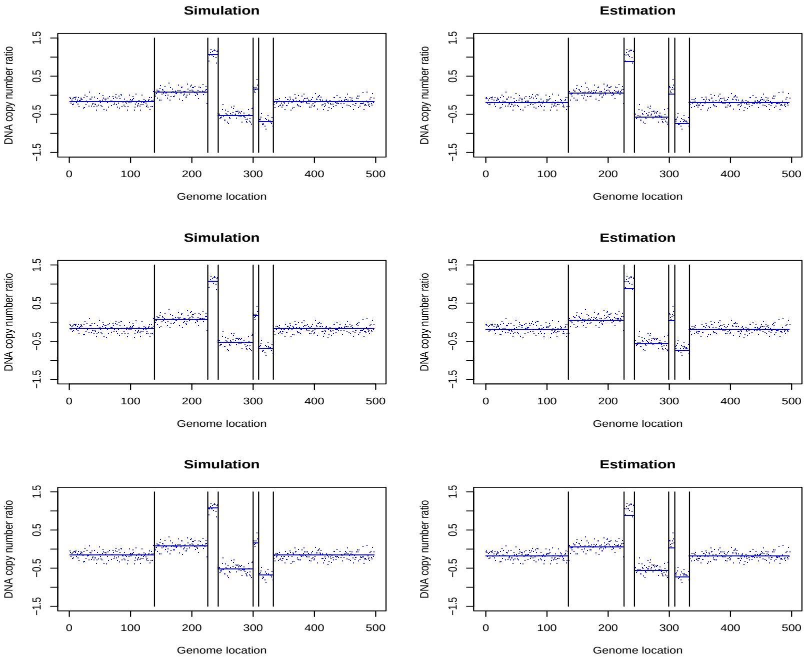

Figure 3: DNA copy number segment estimation ($\sigma = 0.1$). (From top to bottom: no trend, short period and long period; the left panels are simulation set-ups and the right panels are posterior estimations).

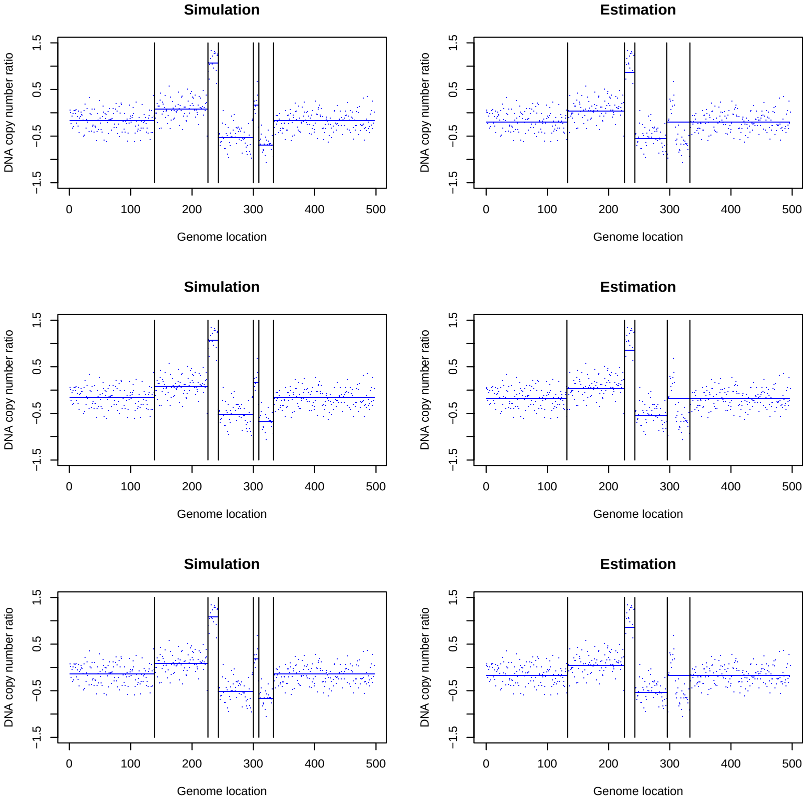

Figure 4: DNA copy number segment estimation ($\sigma = 0.2$). (From top to bottom: no trend, short period and long period; the left panels are simulation set-ups and the right panels are posterior estimations).

$y_{kj} = \mu_k + \epsilon_{kj}$ ($y_{kj} \in$ segment k), where μ_k is the mean and ϵ's are the errors independently distributed as N(0,σ^2). A local trend component was incorporated into the mean for robustness study by $\mu_k = f(k) + 0.25\sigma\sin(a\pi k)$. The normal distribution variation σ is set to be 0.1 or 0.2, and the trend parameter a was set to be 0 (no trend), 0.01 (short period) or 0.025 (long period). For Gibbs sampling, the circle length L is 497 (as a continuous value) and the reference circle origin is fixed between marker 1 and marker 497 after end connection. We assign four move type probabilities: $\pi(H)$=0.05, $\pi(P)$=0.05, $\pi(+)$=0.45, $\pi(-)$=0.45, $\pi(K) \propto$ U[1,100], $\kappa_0 = 1$, $\nu_0 = 12$, σ_0^2=1/12 and μ_0=0.0 (refer to

(2.1)). The starting change point number is K=(1+100)/2=51, equal circular 51-partition creates initial change point (segment) locations, the burn-in is 5,000 and the thinning is 2,000. The results are given in Figures 3 (σ=0.1) and 4 (σ=0.2), where three rows of panels from top to bottom are: no trend, short period and long period respectively. We find that, when the noise has a smaller dispersion (σ=0.1, Figure 3), it recovers all of the original change points, while when the noise has a larger dispersion (σ=0.2, Figure 4), it recovers 5 out of original 6 change points, where certain smaller segments may be combined into one segment. The trend period is not crucial for algorithm performance and equal tails are computationally detected by observing no change point between 1 and 491 circular location (the reference origin on the circle).

4. Discussion

For multiple distributions with same set of regulating parameters (segment starting location t, normal mean μ, and normal variance σ^2, for example) in the circular domain, we make use of an interesting probabilistic process which could be taken as a symmetric transition adjusted by the ratio of differ-by-one segment numbers to implement the posterior parameter sampling. While the algorithm introduced here enjoys substantial simplicity under this special circumstance, we would like to emphasize that, the prior elicitation is crucial for reliable posterior statistical inference, since highly informative and reasonable prior specifications will improve the segment pattern estimation. We anticipate that, the general while simple change point detection by means of Bayesian algorithm developed here will help researchers with biomarker identification within the bioinformatics community.

Acknowledgements

We are grateful to the editor-in-chief, Dr. Min-Te Chao, for his encouraging and insightful suggestions, the associate editor and the referee for very valuable comments which greatly improved our presentation.

References

Barry, D. and Hartigan, J. A. (1992). Product partition models for change point problems. *The Annals of Statistics* **20**, 260-279.

Barry, D. and Hartigan, J. A. (1993). A Bayesian analysis for change point problems. *Journal of the American Statistical Association* **88**, 309-319.

Denison, D. G. T., Holmes, C. C., Mallick, B. K., Smith, A. F. M. (2002). *Bayesian Methods for Nonlinear Classification and Regression.* John Wiley.

Fearnhead, P. (2005). Exact Bayesian curve fitting and signal segmentation. *IEEE Transactions on Signal Processing* **53**, 2160-2166.

Fisher, N. I. and Lee, A. J. (1986). Correlation coefficients for random variables on a unit sphere or hypersphere. *Biometrika* **73**, 159-164.

Fouts, D. L., Wolstenholme, D. L. and Boyer, H. W. (1978). Heterogeneity in sensitivity to cleavage by the restriction endonucleases EcoRI and HindIII of circular kinetoplast DNA molecules of Crithidia acanthocephali. *The Journal of Cell Biology* **79-I**, 329-341.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman & Hall.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711-732.

Huang, T., Wu, B., Lizardi, P. and Zhao, H. (2005). Detection of DNA copy number alterations using penalized least squares regression. *Bioinformatics* **21**, 3811-3817.

Kallioniemi, A., Kallioniemi, O. P., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, F. and Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258**, 818-21.

Lai, Y. and Zhao, H. (2005). A statistical method to detect chromosomal regions with DNA copy number alterations using SNP-array-based CGH data. *Computational Biology and Chemistry* **29**, 47-54.

Liu, J., Yu, W., Wu, B. and Zhao, H. (2006). Bayesian mass spectra peak alignment from mass charge ratios. *Cancer Informatics*, to appear.

Loschi, R. H., Cruz, F. R. B. and Arellano-Valle, R. B. (2005). Multiple change point analysis for the regular exponential family using the product partition model. *Journal of Data Science* **3**, 305-330.

Lucito, R., Healy, J., Alexander, J., Reiner, A., Espositio, D., Chi, M., Rodgers, L., Brady, A., Sebat, J., Troge, J., West, J. A., Rostan, S., Nguyen, K. C., Powers, S., Ye, K. Q., Olshen, A. B., Venkatraman, E., Norton, L. and Wigler, M. (2003). Representational oligonucleotide microarray analysis: a high resolution method to detect genome copy number variation. *Genome Research* **13**, 2291-2305.

Mailfert, A., Vincent-Viry, O. (2001). 2-D inverse problems with applications to microgravity and energy storage. *The International Journal for Computation and Mathematics in Electrical and Electronic Engineering* **20**, 869-878.

Olshen, A. B., Venkatraman, E. S., Lucito, R. and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557-572.

Sen, A. and Srivastava, M. S. (1975). On tests for detecting a change in mean. *Annals of Statistics* **3**, 98-108.

Stanfield, S. W. and Lengyel, J. A. (1979). Small circular DNA of Drosophila melanogaster: chromosomal homology and kinetic complexity. *Proceedings of the National Academy of Sciences of the United States of America* **76**, 6142-6146.

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components - An alternative to reversible jump methods. *The Annals of Statistics* **28**, 40-74.

Yao, Y. (1984). Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches. *Annals of Statistics* **12**, 1434-1447.

Received September 4, 2006; accepted December 9, 2006.

Junfeng Liu
Department of Statistics
Case Western Reserve University
Cleveland, OH 44106, USA
jxl322@case.edu, jfliu@stat.wvu.edu.

E. James Harner
Department of Statistics
West Virginia University
Morgantown, WV 26506, USA
jharner@stat. wvu.edu.

Harry Yang
MedImmune, Inc.
One MedImmune Way
Gaithersburg, MD, 20878, USA
yangh@medimmune.com